Elastic properties of proteins: insight on the folding process and evolutionary selection of native structures

Cristian Micheletti
International School for Advanced Studies (S.I.S.S.A.) and INFM, Via Beirut 2-4, 34014 Trieste, Italy

Gianluca Lattanzi
Hahn-Meitner Institut, Abt. SF5, Glienicker Str. 100, 14109 Berlin, Germany

Amos Maritan
International School for Advanced Studies (S.I.S.S.A.) and INFM, Via Beirut 2-4, 34014 Trieste, Italy
The Abdus Salam International Center for Theoretical Physics, Strada Costiera 11, 34100 Trieste, Italy

(Dated: February 1, 2008)

We carry out a theoretical study of the vibrational and relaxation properties of naturally-occurring proteins with the purpose of characterizing both the folding and equilibrium thermodynamics. By means of a suitable model we provide a full characterization of the spectrum and eigenmodes of vibration at various temperatures by merely exploiting the knowledge of the protein native structure. It is shown that the rate at which perturbations decay at the folding transition correlates well with experimental folding rates. This validation is carried out on a list of about 30 two-state folders. Furthermore, the qualitative analysis of residues mean square displacements (shown to accurately reproduce crystallographic data) provides a reliable and statistically accurate method to identify crucial folding sites/contacts. This novel strategy is validated against clinical data for HIV-1 Protease. Finally, we compare the spectra and eigenmodes of vibration of natural proteins against randomly-generated compact structures and regular random graphs. The comparison reveals a distinctive enhanced flexibility of natural structures accompanied by slow relaxation times at the folding temperature. The fact that these properties are intimately connected to the presence and assembly of secondary motifs hints at the special criteria adopted by evolution in the selection of viable folds.

INTRODUCTION

The backbone structure of globular proteins displays a notable degree of order and organization resulting in secondary motifs such as $\alpha-$helices and $\beta-$sheets and in their optimal arrangement into a compact shape. The special neatness and regularity of native conformations impacts on, or rather reflects, several features that are distinctive of naturally occurring proteins [1, 2, 3, 4]. Indeed, experimental and theoretical studies have shown that native structures conceal a wealth of information about aspects as diverse as native elastic properties, folding rates and key folding stages [5, 6, 7, 8]. A promising graph theoretical approach [9, 10] can be used to ifentify rigid and flexible regions in proteins, and to investigate the emergence of such flexible regions in unfolding processes. Recent topology-based approaches, which are independent of sequence specificity, have revealed that a schematic, coarse-grained structural description is sufficient to obtain results in remarkable qualitative and even quantitative agreement with known experimental facts.

For example, concerning near-native vibrations, one is entitled to replace the detailed interactions among amino acids which are in contact in the native state with springs acting on effective centroids (usually the C_α atoms) [5, 11]. Topological folding models, on the other hand, try to characterize how the loss of configurational entropy contrasts the progressive establishment of native interactions in a folding process [3, 8, 12, 13, 14, 15, 16, 17].

In the present study, we focus on a topology-based model that incorporates both the aspects mentioned above. It consists of a beads-and-springs model but, at variance with other approaches, the strength of the effective interaction depends on the temperature of the thermal bath [18]. Since the model is amenable to analytic treatment, it is possible to characterize rigorously both the thermodynamics as well as the vibrational/relaxation dynamics at various equilibrium stages of the folding process (not only near the native state).

The scope of the present study is two-fold. First we examine how the organization of native contacts impacts on fundamental properties of proteins and, hence, to what extent the knowledge of the contact map can be exploited to predict experimentally verifiable quantities. Our second goal is to repeat the same analysis in the context of generic disordered globular structures. From the comparison of the outcomes we aim at finding clues about the criteria

adopted by nature in the selection of viable folds.

The vibrational spectrum at finite temperature allows to determine typical relaxation times which, in turn, correlate with high accuracy with experimental folding rates. This is rather surprising and unexpected since the relaxation time refers to near equilibrium situations whereas the folding process is far from equilibrium.

Furthermore, the comparative study of analogous static and vibrational properties of disordered globular structures, or otherwise random graphs, and proteins shows that the latter have very uncommon properties in terms of flexibility and mechanical relaxation times that can be exploited to predict not only folding rates but also the key sites of an enzyme and amino acid vibrational amplitudes in partially folded states. These distinctive mechanical properties reflect the hierarchical assembly of protein native states [19] and, in particular, can be traced back to the presence and organization of secondary elements. Hence, they add to the increasing evidence that the emergence of such motifs was promoted by special criteria operated by nature to select viable folds [2, 20, 21, 22, 23].

The straightforward implementation of concepts presented here, makes the model under consideration a useful tool to characterize the folding properties of a protein, complementing alternative theoretical strategies or more labor-intensive experimental techniques.

THEORY

The model under consideration incorporates two features that previous theoretical investigations have proved successful for protein modeling. In spirit, it belongs to the class of Go-models [12], since the energy scoring function introduces a bias in structure space favoring the formation of native contacts. This ensures that the native structure under examination has minimum effective energy. Furthermore, we make use of the observation that, near equilibrium, the dynamics resulting from complicated atomic interactions can be well-reproduced by simple harmonic potentials [11, 24, 25] at least for time intervals shorter that 1 ns [26]. The starting point of our analysis is the following effective harmonic Hamiltonian [18] (for its derivation, see Appendix):

$$H = \sum_{i,j} \mathbf{r_i} L_{ij} \mathbf{r_j} + f(T) \ , \tag{1}$$

where $\mathbf{r_i}$ denotes the distance-vector of site i from its position in the native structure, and f is a term that only depends on the equilibrium temperature. The entries of the symmetric matrix, L, incorporate the elastic forces that act on each residue to restore the native separation with its nearest neighbors in sequence (effect of the peptide bonding) and with other amino acids in interaction in the native state. The peptide bond is modeled by springs whose associated Boltzmann weight is independent of temperature. This reflects the little change of the peptide coupling over the range of temperatures where unfolding/refolding transitions are studied. In addition to this contribution, at variance with previous approaches, the energy scoring-function includes temperature-dependent interactions between the pairs of residues contacting in the native state. The list of such interactions is summarized in the contact map, Δ, whose entries, Δ_{ij}, are 1 if residues i and j are in contact in the native state (i.e. their C_α separation is below the cutoff $c = 7.5$ Å) and 0 otherwise. For simplicity we refer to these temperature-dependent interactions as non-covalent, although they also act on consecutive residues. The strength of such non-covalent springs is calculated self-consistently according to the method described in Appendix . This makes the model non-linear but still very tractable. When the temperature of the model is zero (this mathematically-convenient idealization corresponds to a physiological temperature where the protein is stable in its native state), all springs have the same strength. As temperature is switched on, this strength is reduced by a deterministic amount that is larger for springs where the largest stretching is observed (see Appendix).

RESULTS AND DISCUSSION

Vibrational properties of proteins and disordered structures

At zero temperature, when the non-covalent springs are equally strong and dominate in number over the covalent ones, the model is equivalent to the Gaussian Network Model (GNM) of Bahar *et al.* [27, 28, 29], which has been widely used to characterize the mobile regions of the native state from the analysis of the normal modes of the structure. These represent the natural oscillations (independent of each other) of a protein and convey information about which regions are more flexible, and how thermal energy is dissipated in order to restore equilibrium (i.e. the

native structure). Although atoms in proteins are as tightly packed as in crystalline solids, the vibrational properties of native structures are very different [5]. A notable example is given by the density of eigenvalues of the L matrix, that is the histogram of vibrational frequencies. This is shown in Fig. 1, where we concentrated on two examples: an individual enzyme, the monomer of the HIV-1 protease, and the average histogram calculated for several proteins known to fold via the simplest possible mechanism (two-state) [30]. The peculiarities of the protein spectrum can be ascribed to several distinctive protein properties such as the varying degree of site connectivity (burial profile) or the neat hierarchical organization of the native state in terms of secondary motifs.

To identify how the vibrational spectrum is affected by these general features we shall not take the crystalline solids as a reference, but consider several families of disordered contact maps. For example, to isolate the role played by the burial profile from other features we first consider disordered matrices obtained by taking the contact map of a real protein and randomly reshuffling its entries yet preserving the symmetry of the matrix, the burial profile and the contacts between consecutive sites. Successively we shall consider the case of completely disordered maps but with equal connectivity (burial) for each site. It is important to stress that in these two cases of disorder, the Δ matrices will not, in general, be physically viable. In fact, it is not guaranteed that there exists a viable three-dimensional backbone associated to an arbitrary symmetric contact map. For this reason, we complete our analysis by considering the case of contact matrices of three-dimensional compact self-avoiding structures generated randomly with a computer with the constraint to be as compact as naturally-occurring proteins. This case will provide a useful term of comparison for identifying the spectral properties resulting from the neatly organized three-dimensional shape of naturally occurring protein conformations.

We begin by considering the first instance of disorder, i.e. random reshufflings that preserve both the symmetry of Δ and also the protein native burial profile, i.e. the native number of contacts to which each site takes part [3]. The reshufflings consist in picking randomly a pair of non-zero distinct entries, Δ_{ij} and Δ_{kl} and checking if the entries Δ_{il} and Δ_{kj} are zero. If this is so, the old pair of entries (and their symmetric counterparts) are set to zero and the new ones to 1. The preservation of the connectivity of each site allows to study the impact of disorder on the vibrational frequency spectrum while keeping burial profiles unaltered. As anticipated above, such randomized contact maps are not necessarily the counterpart of any feasible three-dimensional structure; in fact, a more appropriate view of such matrices is within the framework of graphs [31, 32], (the nodes and links corresponding to beads and springs, respectively).

A very important difference shown in Fig. 1 is that proteins have a larger number of low-frequency modes of vibration compared to the spectrum of the reshuffled maps. This difference is made more dramatic by considering the spectrum of loopless regular graphs. The term *regular* here refers to the fact that all sites have the same connectivity (burial), k, but the entries of the associated symmetric matrix, Δ, are otherwise random [33].

This limit case of disorder is important because the average spectrum of such ensemble of matrices has been calculated exactly by McKay [34]. The density of eigenfrequencies is given by:

$$f(\omega) = \begin{cases} \frac{k\sqrt{4(k-1)-(\omega-k)^2}}{2\pi\ (k^2-(\omega-k)^2)} & \text{for } |\omega-k| < 2\sqrt{k-1} \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

and its distribution is shown in Fig. 1 for the case $k = 8$, corresponding to the average site connectivity (including peptide bonding) in real proteins (when the interaction cutoff is around 7.5 Å).

It appears that the range of vibrational frequencies is severely limited compared to that of real proteins; in fact, the spectrum of such graphs reproduces only the central part of the frequency histogram of proteins. The outlying tails at both low and high frequency are not captured at all. Previous studies have pointed out how such tails reflect the existence of heterogeneity in site connectivity (burial) [36, 37]. This fact, and its implications for the mechanical properties of natural biopolymers are examined in detail in the next section.

Localization properties

In a variety of contexts, from the physics of disordered systems to graph theory, spectra have always attracted considerable attention because of the special localization properties of the associated modes. High frequency normal modes are usually concentrated on sites/regions with high connectivity [24, 35, 36, 37, 38, 39]. This is intuitive since a larger number of connections leads to an enhanced local stiffness and hence a higher frequency of vibration. Conversely, low frequency modes are centered on sites/regions that, having fewer connections than average, are more flexible [29, 36]. This overall picture applies remarkably well to the reshuffled contact maps, as visible in Fig. 2. The

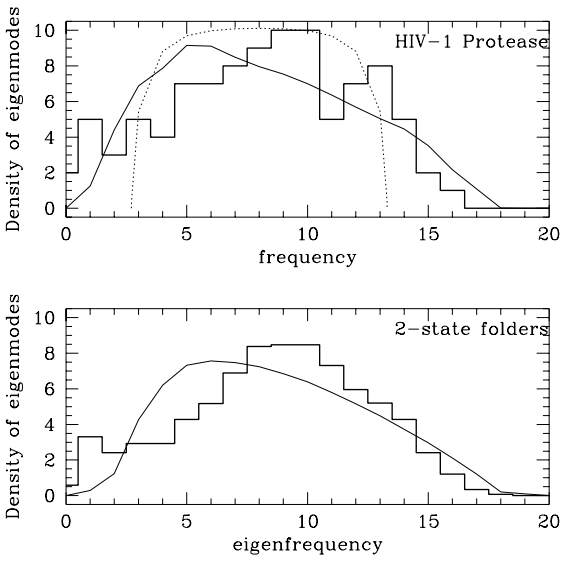


FIG. 1: (top) Density of eigenmodes for (a) HIV-1 Protease (heavy line), (b) average over 3000 conservative reshufflements of HIV-1 PR contact map (light line). The dotted line is the exact density of eigenmodes for a regular graph with $k = 8$ (see text). (bottom) Average density of eigenmodes for the two-state folders listed in the caption of figure 5 (heavy line) and for 100 contact map reshufflements for each of them (light line). Reported data refers to near-native vibrations, corresponding to $T = 0$ in our model. Previous theoretical studies, carried out with full atomic detail, have shown that the peak of the eigenmode distribution falls around a frequency of 45 cm^{-1} [24, 35].

plots show that the modes at low [high] frequency are very localized on sites with low [high] connectivity (burial). The degree of localization of a normalized eigenmode of vibration, $\{v\}$ is defined as $\sum_j |v_j|^4$ (see eg. [37]). This measure will take on the largest value, 1, if all entries of $\{v\}$ but one are zero (full localization). On the contrary, the localization measure will be minimum when all the vector components are equal in modulus (full delocalization).

It is now interesting to turn to the case of proteins. Fig. 2 shows that high-frequency modes are, as before, highly localized on heavily buried sites. Notice that the highly localized states are not limited to the upper edge of the spectrum. The picture, however, changes for low-frequency modes, which appear to be significantly more delocalized than for the reshuffled case. This denotes a degree of flexibility that is significantly larger than randomly connected graphs (yet with the same burial profile!). This is not a peculiar feature of the HIV-1 Protease monomer, but is more general, as illustrated in the right panel of Fig. 2 displaying averages taken over the two-state folders listed in the caption of Figure 5 and their reshufflements. The averages in Fig. 2 are reported only for the 40 slowest modes, which cover a similar range of frequencies across the two-state folders. The inhomogeneity of the lengths of such proteins prevents from taking straightforward averages over the whole frequency range.

Several previous investigations of normal modes and spectra of proteins [24, 25, 35, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49] had already noted the existence and utility of low-frequency modes in proteins. For example the low-frequency spectrum affects not only the mechanical stability of proteins but also the thermodynamic one [50, 51, 52]; furthermore it has been argued that the slowly moving regions are the natural candidates for biological functions involving structural rearrangements. It is important, however, to note that also the burial-preserving reshuffled maps have low-frequency vibrations.

The novel message of Fig. 2 is that the slow modes of random graphs are nearly fully localized (as high frequency modes), while in proteins one encounters the slow movement of regions spanning several residues [24]. This special property, which does not depend on the burial profile, must result from the special organization of native contacts which provides a remarkable degree of flexibility. This, in turn, is associated with a rate of thermal energy dissipation that is slower compared to disordered graphs. The natural measure for the limiting rate at which mechanical excitations decay is, in fact, given by the smallest frequency (eigenvalue) of L. Equivalently, the slowest vibrational relaxation (decay) time in the system, which we shall denote as λ_0, is given by the inverse of the smallest vibrational frequency, $\lambda_0 = 1/\omega_0$.

The immediate cause for this further difference is simply described in terms of graph properties [31]. There is a well-defined relationship between the slowest relaxation time and the average contact-diameter (meaning the average

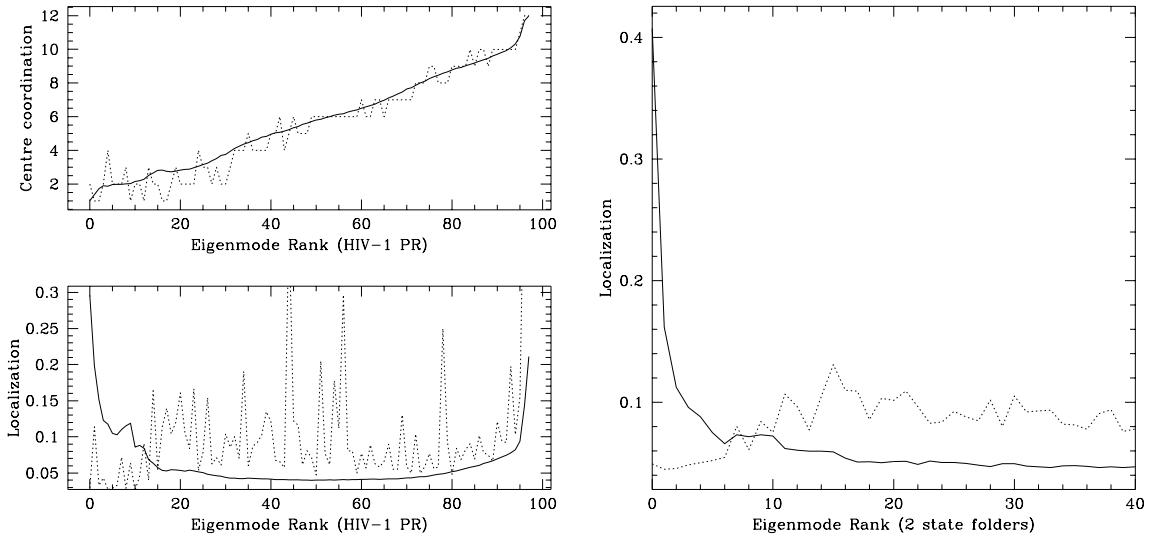


FIG. 2: (top left) Degree of burial (number of native contacts) of sites that are at the center of the various eigenmodes. The eigenmodes are ranked according to increasing vibrational frequency. The center of an eigenmode is the site where one encounters the largest component (in absolute value).(bottom left) Degree of localization of the eigenmodes. The smooth solid lines pertain to averages over 3000 reshufflings of the protease (monomeric) contact map (but preserving the site connectivity). The dotted line relates to the native state. (right) Degree of localization of the first 40 eigenmodes averaged over the two-state-folders (dotted line) and over 100 reshufflements for each of them (solid line).

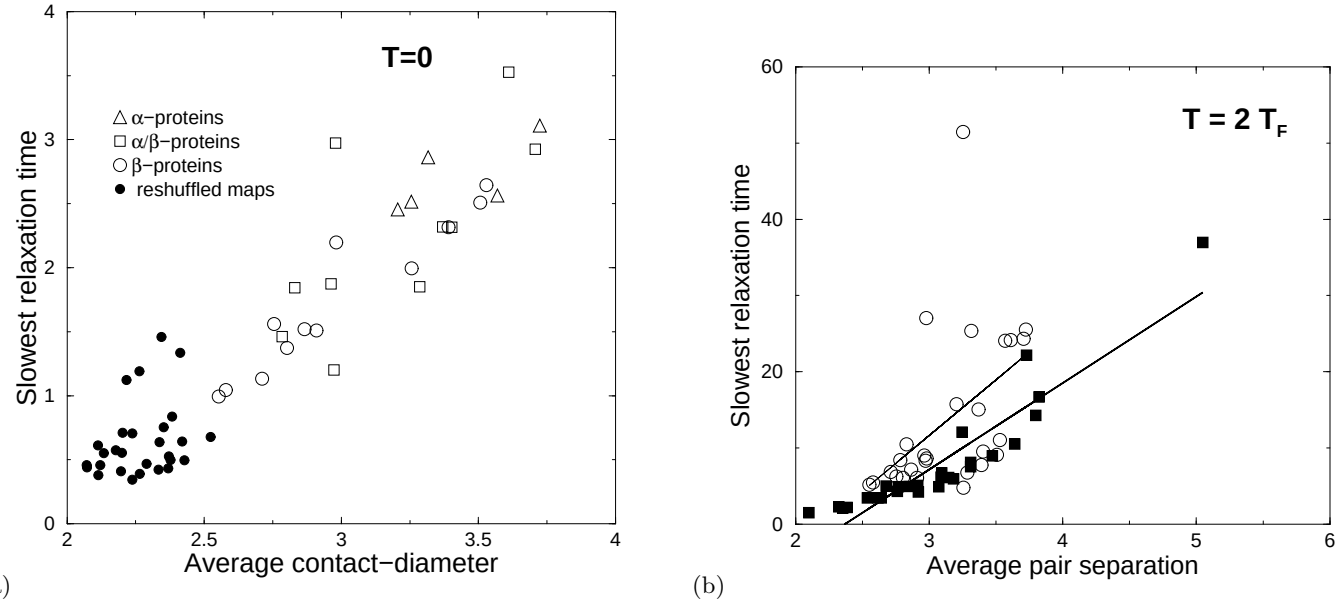
number of native contacts, including those between consecutive sites, that have to be traversed to go from an arbitrary site to another). This relation is neatly visible in Fig. 3, which displays results for the distinct two-state folders and an equal number of their reshufflements.

It is important to see that, among all proteins, the ones with many local contacts (all-α family) tend to be at one end of the scatter plot, while proteins belonging to the all-β family are closer to the case of reshuffled contact maps. The relaxation of real proteins appears to be also slower than randomly-collapsed three-dimensional structures. Within our simplified approach this difference is most visible at finite temperatures, as illustrated in Fig. 3, where the relaxation rates are computed at $T = 2\,T_F$.

The slowest relaxation encountered in proteins can be understood from the fact that the establishment of a few random contacts, very rapidly makes the structure rigid, and less susceptible to further rearrangements. The hierarchical organization of contacts in proteins, in terms of secondary motifs further arranged in the tertiary structure, is responsible for the enhanced flexibility visible in Fig. 3. In fact, the relaxation times of *isolated* secondary motifs, both α and β, of equal length is nearly the same and, again, much higher than a corresponding random reshuffling. This is illustrated in Fig. 4. From these facts it is tempting to speculate that α- and β-like motifs are among the configurations that, for a given burial profile, have the largest relaxation times. This is particularly intuitive in the case of helices since their periodicity (modulus boundary effects) is immediately associated to the presence of slowly-decaying excitations.

Folding rates

In this subsection we shall explore the intimate connection between the relaxation time of slow eigenmodes of protein structures in thermal equilibrium and the protein folding rate. It is tempting, and physically appealing, to speculate that the relaxation rate of partially folded states may be an indicator of the folding velocity. This stems from the observation that low-frequency vibrations involve extended protein segments and require less energy to be activated. Hence, structural rearrangements from the unfolded state to the native one should occur more easily when the relaxation times of partially folded states are higher. This qualitative argument is also supported by a different reasoning where the folding of a protein is viewed as a diffusion in the complicated free-energy landscape until the native-state (at the minimum of the accessible free energy) is reached. Different native shapes will then be associated with landscapes that, despite being biased towards the native conformation, may have significant differences in terms



(a) (b)

FIG. 3: (left) Scatter plot of the slowest native relaxation times ($T = 0$) versus the average native contact diameter. The open symbols denote data pertaining to the two-state folders listed in the caption of Fig. 5. The open triangles, square and circles denote proteins belonging to the α, β and $\alpha\beta$ families respectively. The filled circles denote the slowest relaxation time observed with a random (but burial-preserving) reshufflement of the native contact map of each of the two-state folders. (right) Scatter plot of the slowest relaxation times at $T = 2\,T_F$ versus the average native contact diameter. The data pertain to the two-state folders and an equal number of computer-generated disordered compact structures with same length distribution (filled squares).

of numbers of metastable minima, height of the barriers between them etc. The corrugation of the landscape will undoubtedly affect the folding rate. The relaxation time of a structure in thermal equilibrium reflects the average ruggedness of the landscape. In a smoother landscape a protein will be able to make larger conformational changes for the same amount of dissipated energy. We can thus conclude that the native state is reached with greater difficulty if, during the folding process, the partially-formed structures have smaller equilibrium relaxation times.

In agreement with this picture, we have found a notable interdependence (see Fig. 5) between the experimental folding rates, K_F, and the slowest relaxation rate, ω_0 (in dimensionless units) which dominates the long-time relaxation kinetics in equilibrium. At temperatures much larger than the folding temperature, T_F (identified as the temperature at which the peak of the specific heat is observed), the relaxation time, $\lambda_0 = 1/\omega_0$ conveys little information about the folding rate, since virtually no contacts are formed; this is true also at low T, where it measures only the rate of thermal dissipation of the fully-folded native structure.

The highest correlation is observed at finite temperatures higher than T_F (see Fig. 5). Although we believe that the precise location of the peak of the correlation may be sensitive to the model details, this result is plausible since it is above the folding transition that the significant search in conformation space occurs. The correlation coefficient in the neighborhood of the optimal temperature is $r = 0.73$. It is possible to assess its statistical significance by comparison with the null case of no correlation between relaxation rates and folding times. In the absence of any pair correlation between variables with converging moments, one expects that the correlation coefficient measured over a finite sample is normally distributed. This allows to calculate exactly the probability to encounter a correlation coefficient higher than $r = 0.73$ if relaxation and folding times were statistically independent. This probability turns out to be equal to $2 \cdot 10^{-4}$ which testifies that the observed dependence between the two quantities is truly significant.

These results add to previous studies where folding rates were predicted with analogous level of significance on the basis of contact locality [6] or different topologic indicators (such as the clustering coefficient or cliquishness) based on a graph-theoretical description of a protein's non-bonded interactions [32].

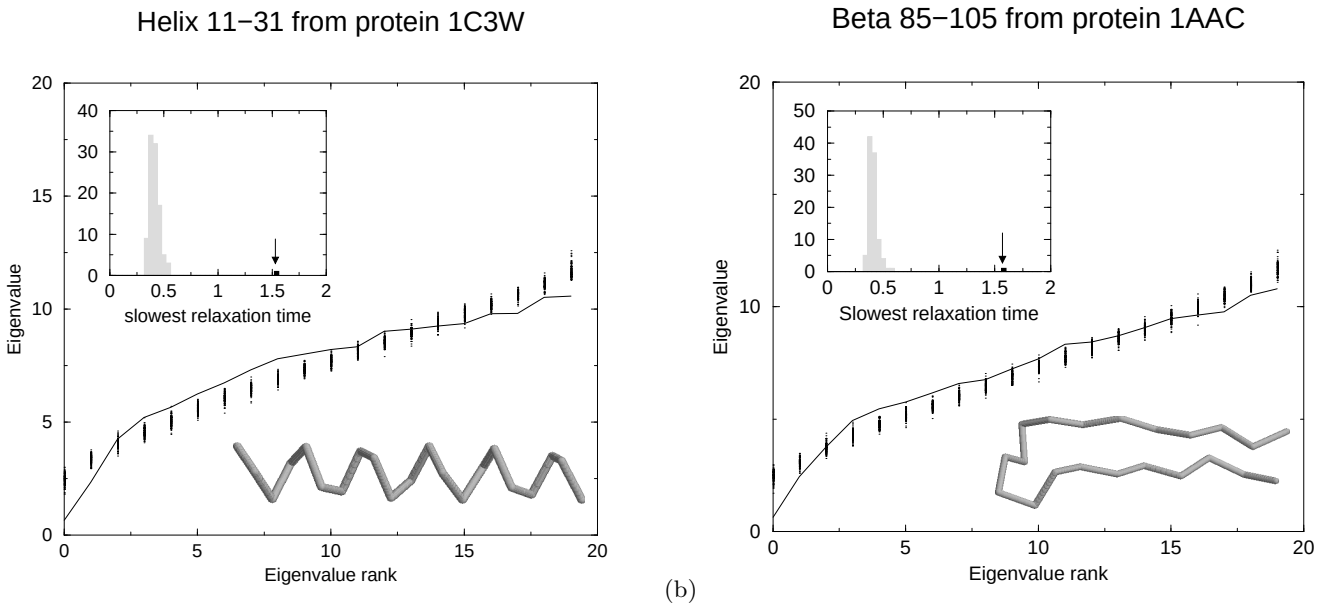


FIG. 4: Frequency spectrum at $T = 0$ of (a) helix 11-31 from protein 1c3w and (b) the beta sheet 85-105 from protein 1aac (continuous line). The dots denote the frequency spectrum calculated over 100 burial-preserving random reshufflements of the original contact map. The inset displays the histogram of the slowest relaxation times, λ_0 for the reshuffled maps. The slowest relaxation time for the original helix and beta sheet are denoted by an arrow in the respective figures.

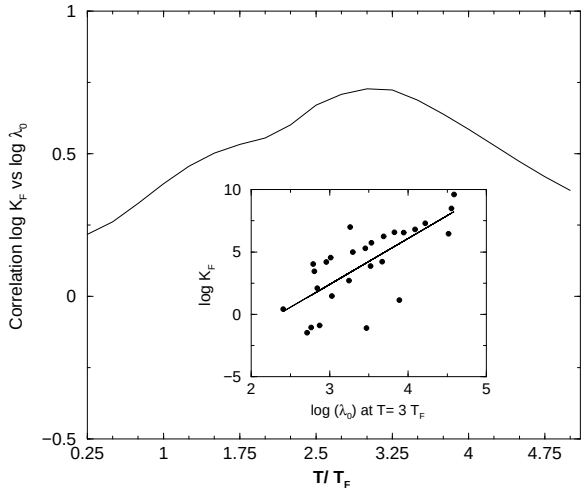


FIG. 5: Linear correlation coefficient between the logarithm of the theoretical relaxation time, $\log \lambda_0$, and experimental folding rates for a set of proteins known to fold via a two-state mechanism. The inset shows the scatter plot of the actual theoretical and experimental values at a temperature close to where the best correlation, $r = 0.73$ is observed. Notice that K_F spans several orders of magnitude. The folding rates were desumed from the experimental studies collected in [6, 30, 32, 53, 54, 55]. The set contained the proteins listed below. α family: 1lmb, 2abd, 1imq, 1ycc, 1hrc; α/β family: 2gb1, 1div.n,2ptl, 1coa, 1hdn, 1div.c,1urn, 1aps, 1fkb, 2vik; β family: 1shg,1srl,1shf.a, 1tud,1csp,3mef,2ait,1pks,1ten,1wit,1fnf (9FN-III and 10FN-III). When multiple experimental entries were available for the folding rate of a give protein, we used considered the arithmetic average of $\log(K_F)$.

B factors

The choice of a simple Gaussian Network Model to investigate the biological function of proteins has been motivated by a striking qualitative agreement between the experimental B factors (or temperature factors) measured in X-ray diffraction experiments and the mean square residue displacement calculated theoretically from the expression [28, 29, 56]:

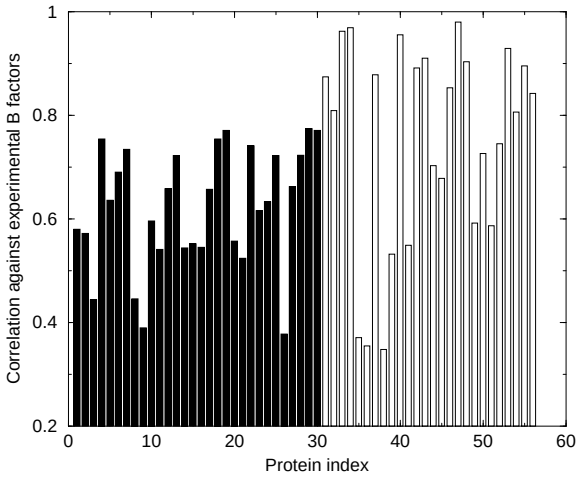


FIG. 6: Correlation between the theoretical B factors and the experimental ones reported in 30 X-ray resolved proteins (filled bars) and NMR resolved ones (open bars). The proteins in the two classes were ordered according to increasing lengths (the overall length range is 39-258). The PDB code of the proteins adopted here are reported below. **X-ray** (1 to 30):1shg, 1fxd, 1igd, 1ail, 1hoe, 1vcc, 1cei, 1opd, 1fna, 1pdr, 1beo, 1poa, 1mai, 1bfg, 1laa, 1kuh, 1cof, 1lcl, 1pkp, 1vsd, 1npk, 1vhh, 1gpr, 1sfe, 1amm, 1ido, 1kid, 1cex, 1chd, 1thv; **NMR** (31 to 56): 1aqr, 1aa3, 1ikl, 1bo0, 1awj, 1a8c, 1a1w, 1bhu, 1ocd, 1ght, 1trs, 1jli, 1fwq, 1tam, 1ahk, 1tbd, 1gdf, 1joo, 1fls, 1buy, 1il6, 1aa9, 1ak6, 1a23, 1lxl, 1eza.

$$B_i^{\rm theory} \propto \sum_k \frac{1}{\omega_k} |v_i^k|^2 \tag{3}$$

where i is the residue index, v_i^k is the i-th component of the k-th mode whose eigenfrequency is ω_k. The sum in eqn. (3) is over k such that $\omega_k \neq 0$. Several alternative methods exist to predict (or refine) experimental B factors [24, 35, 57, 58, 59]. The appealing feature of this Gaussian approach is that it is possible to give a good account of the experimental B factors by exploiting only the connectivity information contained in the native structure. A recent work by Halle [59] has remarked the heavy influence that the native-state topology exerts on the measured B-factors; in particular, it was shown that the mobility of an amino acid anti-correlates with its degree of burial (estimated as the number of non-covalent bonds to which it takes part). Besides the fundamental effect of the burial profile, the GNM allows to model the influece on B-factors of other important topogical features, such as the locality of contacts and presence of highly-interconnected clusters of interactions. In addition, the decomposition of dynamical motion into independent modes allows to trace which modes of vibrations are most responsible for the mobility of a given site or protein region, thus providing useful hints about the protein biological function.

In this section we perform an analysis of the correlation between the theoretical and experimental B factors, aimed at investigating the role played by the single vibrational eigenmodes in the overall observed temperature factors.

We stress that in our model, both eigenfrequencies and eigenmodes are temperature-dependent. We begin our analysis by considering the $T \to 0$ limit of our model, that corresponds to the GNM studied in refs. [27, 28, 44].

We analyzed 30 protein structures, all obtained in X–ray diffraction experiments, of different size (up to about 200 amino acids). The proteins, listed in the caption of Fig. 6, are chosen among the monomeric representatives of distinct structural classes obtained from high-resolution data [60]. Thermal fluctuations for each amino acid were compared with the corresponding temperature factor reported for the CA atoms in the PDB file [61]. Although a remarkable agreement with experimental data can be obtained for several proteins (correlation coefficients larger than 0.75 for proteins with nearly 200 residues), in some rare instances the correlation was around 0.4 (see Fig. 6). It must be borne in mind, however, that such correlations are never trivial, since they are measured over the entire protein length. Assuming, as before, that in case of no correlation the distribution of r is normal, we can calculate the probability to observe higher correlations than the measured ones. Even in the worst case, this probability is never larger than $t = 3\ 10^{-4}$ which is an indicator of excellent correlation. This estimate of statistical significance is so small that even if more precise calculations (e.g. accounting for the skewness in the B-factors distributions) modify the estimate of t by two orders of magnitude, the correlation significance would be high.

Although statistically relevant, the agreement with temperature factors does not have the same quality for all crystal structures. It is important to remark that the crystallographic B-factors are not obtained by direct measurement,

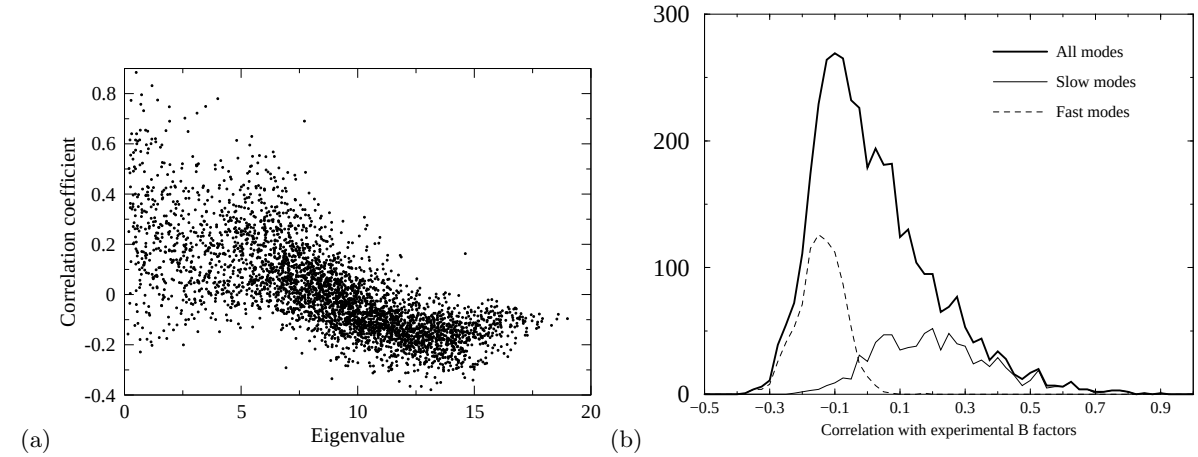


FIG. 7: (a) Correlation coefficients between single eigenmodes amplitudes (only one term at a time is considered in eqn. (3)) and experimental B factors as a function of their corresponding eigenfrequencies. The points refer to all eigenmodes of the X-ray structures listed in the caption of Fig. 6.(b) Histograms from data in (a). Positive correlations come from slow modes, negative from fast modes.

but from a suitable parametric fit of the measured diffraction pattern. In unfavorable conditions (affected also by crystallographic resolution), the fit may be significantly underdetermined and reliable data can be obtained only imposing additional constraints on the B factors (such as their smooth variation along the main- or side-chains). Inspection of the PDB files for which the worst correlations are seen have revealed that, indeed, those proteins lacked the B-factor smoothness. Another possible source of discrepancy between theoretical and experimental data is the fact that we neglected the proximity effect of the surrounding proteins in the crystal structure. The presence of neighboring proteins might reduce the mobility of some regions (and notably those experiencing a pronounced freedom of movement), as already noted in previous studies [24, 35, 58, 59].

Equation (3) clearly indicates that most of the contribution to the theoretical B factor comes from low-frequency modes, since the fast ones are suppressed by their frequency reciprocal weight. It is sensible to ask how much each *individual* mode correlates with the experimental B factors. If a clear trend of such correlation against the mode frequency is found, one could devise a better (knowledge-based) weighting scheme, alternative to the one of eqn. 3 to improve the theoretical estimates of mean square displacement of residues. To answer this question we have obtained for each mode, k, the linear correlations between the experimental B factors along the chain, $\{B_i^{exp}\}$ and the mode amplitudes, $\{|v_i^k|^2\}$. Such correlations, computed separately for each mode of the proteins listed in Fig. 6 are reported in Fig. 7a.

The majority of positive correlations are found in the region of slowest eigenmodes (eigenfrequencies up to 6), while there is a negligible correlation for middle frequencies, and a weak anti-correlation for fast eigenmodes (frequencies larger than 12). This is evident in figure 7b where one can compare the histograms of the correlation coefficients cumulated (with equal weights) over fast and slow eigenmodes. The anti-correlation of the experimental B-factors with the square-amplitudes profile of fast eigenmodes is in agreement with the graph theoretical prediction of a localization of the fastest eigenmodes in nodes of higher connectivity. These nodes are less mobile in the protein structure and hence associated with smaller B values. Furthermore, since all eigenmodes are orthogonal, the sites whose vibrational amplitude is large in the slowest eigenmodes, will not be very mobile in the fastest ones, and viceversa.

Following this analysis we can conclude that low eigenmodes cooperate to generate much of the fluctuation pattern of the protein structure, while intermediate modes have little influence and fast modes could be safely excluded from the sum, since they contribute to raise the mobility of highly connected nodes.

Finally, we observe that the theoretical B's carry a non trivial dependence on temperature through the eigenmodes (due to the temperature dependence of matrix L). Fig. 8 summarizes the behaviour of the correlation with experimental data against temperature for four representative proteins. It can be seen that, typically, by working at a temperature around $2\,T_F$ the correlation with experimental data can be increased significantly. As visible, an exception to this trend is given by the few proteins which already have a poor correlation at $T = 0$.
We have considered also the case of NMR proteins, for which accurate relaxation measurements can directly probe the mobility of local portions of the protein thus providing an *effective* B factor. Strikingly, the correlation between

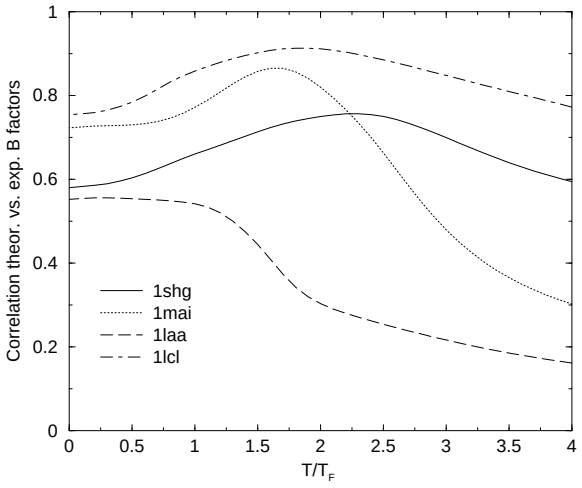


FIG. 8: Correlation between theoretical and experimental B factors as a function of temperature.

the theoretical and the experimental B factors observed in this situation was systematically higher than for the X-ray resolved counterparts, as can be ascertained from Fig. 6. It would be tempting to conclude that this improved agreement is due to the more direct way with which NMR experiments can address local vibrational amplitudes. However, due to the insufficient annotation of PDB files of some NMR-resolved structures, it is not always possible to ascertain whether the reported data follow from precise relaxation measurements or from an *a posteriori* analysis of the various model structures compatible with measured distance-constraints.

B factors and key sites

The analysis of the mobility, in thermal equilibrium, of different protein residues carries significant information about sites that are important in the folding process. We illustrate this important application for the case of the HIV-1 Protease. This enzyme represents an ideal benchmark due to the vast number of clinical studies thanks to which comprehensive tables of key mutating sites have been compiled [62, 63, 64, 65, 66, 67]. Molecular dynamics studies have attempted to clarify the special role played by these amino acids either through all-atom simulations of the protease [68] or through topological folding models [8]. It appears that a number of key mutating sites form contacts that act as rate limiting steps for the folding process [8].

One attempt to describe/predict this crucial set of amino acids from the analysis of normal modes was carried out in Refs. [27, 28, 44] within GNM (to which our model reduces when $T = 0$), who pointed out that high frequency modes are localized close to key mutating sites [45]. Our results, especially those of Fig. 2, support the view that the sites where the high frequency modes concentrate are paramount for native stability; in fact they are among the least exposed ones. However, this property is only related to the native structure, and does not imply that the same sites play the leading role in overcoming the folding rate-limiting steps.

A more direct strategy would be to determine the handful of contacts that form cooperatively at, or in the neighborhood of, the folding transition temperature. The formation of such contacts, which has an all-or-none character, is expected to constrain significantly the mobility of the residues involved in their establishment. This scheme, which has been confirmed *a posteriori* in simplified folding models of the HIV-1 PR and prion [8, 69], can be naturally adopted in the present case. In fact, a natural and convenient measure of the degree of spatial constraint of a given site is provided by the associated B-factor. Thus, near the folding transition temperature, one expects that the key sites have nearly native like values for the B factors, while other residues will have much larger fluctuations than in the native state. Therefore, one may identify the key sites as those with small values of B or small derivatives of B with respect to temperature. It turns out that both these criteria work well, as illustrated in Fig. 9. Among the top 13 sites ranked according to the smallness of B factors there are 5 key residues: 32,33,77,84,30. In the top 13 sites ranked according to the derivative of B there is an additional key site, namely residue 82. The probability to observe at least these many matches had we chosen the sites randomly, would have been 3 % and 0.5 % respectively in the two cases. It must be stressed that the present strategy to identify key sites exploits the information on native topology in a non-trivial manner avoiding, for example, the pitfall of selecting the most buried sites as the key ones. In fact,

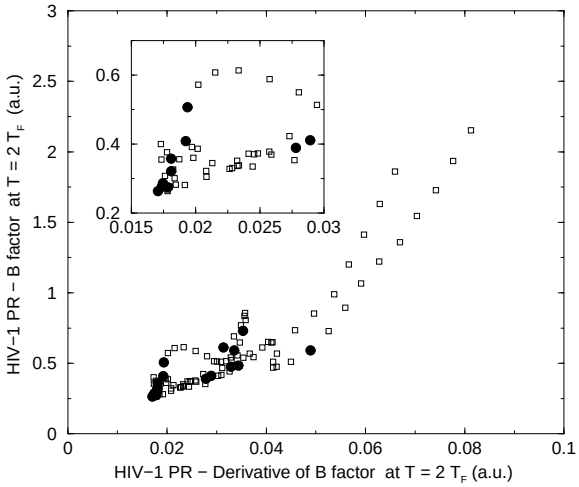


FIG. 9: Scatter plot of B-factors and their temperature derivative for the sites of HIV-1 PR monomer. The key sites known to cause resistance to protease-inhibiting drugs are highlighted.

to collect the same number of correct matches found before, one needs to retain about twice as many burial-ranked sites than for the B-factor-ranked case. We have verified that this conclusion holds for all interaction cutoffs in the range 6.5-8 Å. This very good statistical validation shows that the physically appealing identification of the key sites through normal modes analysis can be fruitfully applied in contexts of high practical importance.

CONCLUSIONS

We have focused on the characterization of both the folding process and equilibrium physical properties of natural proteins through the normal modes analysis of a suitable topological model. Previous investigations of normal vibrations around the native state have proved a valuable tool to describe the large scale motion of proteins and obtain clues about their biological functionality.

Although we also focus on specific protein instances, such as the HIV-1 protease, our scope is more directed at finding some general characteristics of proteins' dynamics in thermal equilibrium (at various temperatures) that distinguish these biopolymers from a randomly compact polymer. This comparison has a two-fold objective: one of theoretical interest, the other more practical.

¿From the point of theoretical biophysics, the identification of features that are unique (and common) to proteins provides clues about the special evolutionary criteria that have promoted the selection of the wide repertoire of naturally-occurring proteins yet using only a limited number of fold families [4, 20, 21]. In the detailed account provided here we have shown that proteins are very flexible and have long relaxation times, in thermal equilibrium, compared to globular polymers. In turn, the relaxation time of partially folded states is shown to have a significant correlation with experimental folding times (r can be as high as 0.73). We give a quantitative account of the intimate dependence of the relaxation rates with native state topology; thereby providing a novel additional support to the influence of native-state topology over folding rate. It is also found that high-frequency modes involve as many sites in natural proteins as in disordered structures. Despite this, the structural regions involved in slow motions are much more extended in proteins. Without this property (common to the few tens of proteins considered here) it would probably be impossible for a protein to carry out the articulated mechanical tasks necessary for biological functionality [45, 46, 47, 48]. The direct investigation of the vibrational properties of individual secondary motifs has shown that their presence and organization are distinctive features of protein structures. This observation adds to previous evidence according to which the ubiquity of secondary motifs in real proteins is due to the very special properties that they confer both to the native state and to the folding process.

¿From a more practical point of view, the knowledge of how these "special features" have arisen through the selection of certain viable native shapes, can be used as a predictive tool. In particular, the straightforward deterministic analysis of the topologic folding model adopted here allows to predict folding rates with high statistical significance. A further important application is in the "a priori" determination of the B factors that are essential in the refinement and validation of protein structures resolved by X-ray and NMR. A by-product of the analysis of B factors of partially

folded states is the identification of the sites (or contacts) that lock into their relative native position at early stages of the folding process. Taking the case of HIV-1 PR as a reference, we have shown that such sites are among those that are known to be involved in rate-limiting steps of the folding process. Therefore, within this single simple framework it is possible to model various aspects of the folding process and other equilibrium properties obtaining a thorough protein characterization that is consistent with available experimental results.

This work was supported by INFM and MURST cofin 2001. GL's research has been supported by a Marie Curie Fellowship of the European Community Programme "Improving Human Research Potential and the Socio–Economic Knowledge Base" under contract number HPMF–CT–2001–01432. The authors are solely responsible for information communicated and the European Commission is not responsible for any view or results expressed. We are indebted to Jayanth Banavar, Ken Dill, Burak Erman, Doriano Lamba and Henriette Molinari for stimulating discussions and helpful suggestions.

APPENDIX. METHODS: ANALYTICAL CHARACTERIZATION OF THE MODEL

The simplified energy functional used to characterize the equilibrium or folding properties of a target structure, Γ is:

$$\mathcal{H} = \frac{K \cdot T}{2} \sum_{i=1}^{N-1} (\mathbf{r}_i - \mathbf{r}_{i+1})^2 + \frac{1}{2} \sum_{i \neq j} \Delta_{ij}^{\Gamma} [(\mathbf{r}_i - \mathbf{r}_j)^2 - R^2] \, p_{ij} \tag{4}$$

where $\mathbf{r}_i$ denotes the distance-vector of the ith C_α from its position in the native structure, Γ (assumed to be N residues long). K is the strength of the peptide bonds, and T is the absolute temperature (incorporating the Boltzmann constant). Δ^{Γ} is the contact matrix, whose element Δ_{ij} is 1 if residues i and j are in contact in Γ (i.e. their C_α separation is below a cutoff c. (in this study $c = 7.5$ Å) and 0 otherwise. The matrix Δ_{ij}^{Γ} along with the native C_α coordinates encodes the topology of the protein. In standard off-lattice approaches, the interaction $V(d)$ between non-bonded amino acids at a distance d, is taken to be a square well potential, or some type of Lennard-Jones interaction. Our choice in Eq. (4) is a sort of "harmonic well" which, while being physically sound and viable, is suitable for a self-consistent treatment, as explained below.

The temperature-dependent term p_{ij} is the probability that the separation of amino acids i and j fall within a distance R of the native separation. Therefore R plays the role of an effective outer rim of the quadratic potential well and can be set to a few Angstroms ($R = 3$ Å in the present study) to reflect the fact that, when the separation of two residues exceeds substantially the native one their interaction is negligible. The strength of the peptide bond, K, is set to 1/15 [18]. Such choice of values guarantees that close to the folding transition temperature, nearly half of the native contacts are formed, consistently with several unrelated studies [30, 70].

The simple quadratic form of Hamiltonian in Eq. (4) allows to determine exactly the p_{ij}'s from the self-consistent relation:

$$p_{ij} = \langle \Theta(|\mathbf{r}_i - \mathbf{r}_j|^2 - R^2) \rangle. \tag{5}$$

where $\Theta(x)$ is the unitary step function and the brackets denote the thermal averaging under the action of Hamiltonian 4, that depends on p_{ij} as well. Operatively, one starts from a trial choice of the p_{ij}, which is used to determine a new set of parameters through eqn. (5). Convergence is obtained in a few tens of iterations at any temperature. Now in such self consistent approach the problem is fully solved and all equilibrium or dynamical quantities can be calculated exactly by evaluating analytical expressions. At $T = 0$ the model correctly assigns the lowest energy to the native structure (all p_{ij}'s equal to 1). Upon increasing the temperature, each interacting pair will have increasing mutual separation (in modulus) and correspondingly, the p_{ij}'s will decrease to reflect the milder binding. In the limit $T \to \infty$ all p_{ij}'s tend to 0. It is worth pointing out that our energy functional treats non-native contacts in a neutral way: their formation is not favoured but is not discouraged either. As a consequence, at finite temperatures, non-native contacts have a non-zero probability of formation, which can still be computed through 5.

Simple algebraic operations allow to recast the self-consistent energy function in the following form:

$$\mathcal{H} = \sum_{ij} \mathbf{r_i} L_{ij} \mathbf{r_j} - \frac{1}{2} \sum_{ij} \Delta_{ij}^{\Gamma} \, R^2 \, p_{ij} \tag{6}$$

where

$$L_{i,j} = \begin{cases} K \cdot T\,(2 - \delta_{i,1} - \delta_{i,N})/2 + \sum_l \Delta^{\Gamma}_{i,l} p_{i,l} & \text{for } i = j \\ -p_{i,j}\Delta_{i,j} + K \cdot T\,\left(-\delta_{i,j+1} - \delta_{i,j-1}\right)/2 & \text{for } i \neq j\ . \end{cases} \tag{7}$$

While the present form of the model does not accurately describe the effects of self-avoidance this does not lead to a qualitatively wrong behaviour in the highly-denatured ensemble (large T). The treatment of steric effects becomes progressively more accurate as temperature is lowered. In fact, the model guarantees that the native state is the true ground state and therefore protein conformations found at low temperature inherit the native self-avoidance. The connectedness of the chain, as well as its entropy, are captured in a simple but non-trivial manner. The most significant advantage of the model is that it can be used to explore the equilibrium thermodynamics without being hampered by inaccurate or sluggish dynamics.

[1] Baker, D. A. (2000). Surprising simplicity to protein folding. *Nature,* **405**, 39–42.
[2] Maritan, A., Micheletti, C., Trovato, A. & Banavar, J. R. (2000). Optimal shapes of compact strings. *Nature,* **406**, 287.
[3] Micheletti, C., Banavar, J. R., Maritan, A. & Seno, F. (1999). Protein structures and optimal folding from a geometrical variational principle. *Phys. Rev. Lett.* **82**, 3372–3375.
[4] Banavar, J., Maritan, A., Micheletti, C. & Trovato, A. (2002). Geometry and physics of proteins. *Proteins: Structure Function and Genetics,* **47**, 315–322.
[5] ben Avraham, D. (1993). Vibrational normal–mode spectrum of globular proteins. *Physical Review B,* **47**, 14559–14560.
[6] Plaxco, K. W., Simons, K. T. & Baker, D. (1998). Contact order and transitian state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**, 985–994.
[7] Chan, H. S. (1998). Matching speed and locality. *Nature,* **392**, 761–762.
[8] Cecconi, F., Micheletti, C., Carloni, P. & Maritan, A. (2001). Molecular dynamics studies of HIV-1 protease: drug resistence and folding pathways. *Proteins: Structure Function and Genetics,* **43**, 365–372.
[9] Jacobs, D. J., Rader, A. J., Kuhn, L. A. & Thorpe, M. F. (2001). Protein flexibility predictions using graph theory. *Proteins: Structure Function and Genetics,* **44**, 150–165.
[10] Rader, A. J., Heispenheide, B. M., Kuhn, L. A. & Thorpe, M. F. (2002). Protein unfolding: rigidity lost. *Proc. Natl. Acad. Sci. USA,* **99**, 3540–3545.
[11] Tirion, M. M. (1996). Large amplitude elastic motions in proteins from a single–parameter, atomic analysis. *Physical Review Letters,* **77**, 1905–1908.
[12] Go, N. & Scheraga, H. A. (1976). On the use of classical statistical mechanics in the treatment of polymer chain conformations. *Macromolecules,* **9**, 535–542.
[13] Galzitskaya, O. V. & Finkelstein, A. V. (1999). A theoretical search for folding/unfolding nuclei in 3D protein structure. *Proc. Natl. Acad. Sci. USA,* **96**, 11299–11304.
[14] Munoz, V., Henry, E. R., Hofrichter, J. & Eaton, W. A. (1999). A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl. Acad. Sci. USA,* **95**, 5872.
[15] Alm, E. & Baker, D. (1999). Prediction of protein folding mechanisms from free energy landscapes derived from native structures. *Proc. Natl. Acad. Sci. USA,* **96**, 11305–11310.
[16] Klimov, D. K. & Thirumalai, D. (2000). Native topology determines force-induced unfolding pathways in globular proteins. *Proc. Natnl. Acad. Sci. USA,* **97**, 7254–7259.
[17] Clementi, C., Nymeyer, H. & Onuchic, J. N. (2000). Topological and energetic factors: what determines the structural details of the transition state ensemble and 'en-route' intermediates for protein folding? an investigation for small globular proteins. *J. Mol. Biol.* **298**, 937–953.
[18] Micheletti, C., Banavar, J. & Maritan, A. (2001). Conformations of proteins in equilibrium. *Phys. Rev. Lett.* **87**, DOI:088102–1.
[19] Baldwin, R. L. & Rose, G. D. (1999). Is protein folding hierarchic? I. local structure and peptide folding. *Trends in Biochemical Sciences,* **24**, 26–33.
[20] Denton, M. & Marshall, C. (2001). Laws of form revisited. *Nature,* **410**, 417.
[21] Chothia, C. (1984). Principles that determine the structure of proteins. *Annu. Rev. Biochem.* **53**, 537–572.
[22] Chan, H. S. & Dill, K. A. (1990). The effects of internal constraints on the configurations of chain molecules. *J. Chem. Phys.* **92**, 3118–3135.
[23] Socci, N. D., Bialek, W. S., & Onuchic, J. N. (1994). Properties and origins of protein secondary structures. *Phys. Rev. E,* **49**, 3440–3443.
[24] Levitt, M., Sander, C. & Stern, P. S. (1985). Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme. *J. Mol. Biol.* **181**, 423–447.
[25] Horiuchi, T. & Go, N. (1991). Projection of monte carlo and molecular dynamics trajectories onto the normal mode axes: human lysozyme. *Proteins: Structure Function and Genetics,* **10**, 106–116.

[26] Amadei, A., Linssen, A. B. M. & Berendsen, H. J. C. (1993). Essential dynamics of proteins. *Proteins: Structure Function and Genetics,* **17**, 412–425.

[27] Bahar, I., Atilgan, A. R. & Erman, B. (1997). Direct evaluation of thermal fluctuations in proteins using a single parameter harmonic potential. *Folding and Design,* **2**, 173–181.

[28] Haliloglu, T., Bahar, I. & Erman, B. (1997). Gaussian dynamics of folded proteins. *Physical Review Letters,* **79**, 3090–3093.

[29] Bahar, I., Atilgan, A. R., Demirel, M. C. & Erman, B. (1998). Vibrational dynamics of folded proteins: significance of slow and fast motions in relation to function and stability. *Physical Review Letters,* **80**, 2733–2736.

[30] Jackson, S. E. (1998). How do small single-domain proteins fold? *Folding and Design,* **3**, R81–R91.

[31] Bollobas, B. (1985). *Random Graphs*. Academic, London.

[32] Micheletti, C. (2001). Optimal prediction of folding rates and transition state placement from native state geometry. *SISSA preprint,* , http://xxx.lanl.gov/ps/cond–mat/0202090.

[33] Merris, R. (1994). Laplacian matrices of graphs: a survey. *Lin. Alg. Appl.* **197**, 143–176.

[34] McKay, B. D. (1981). The expected eigenvalue distribution of a large regular graph. *Lin. Alg. Appl.* **40**, 203–216.

[35] Tirion, M. M. & ben Avraham, D. (1993). Normal mode analysis of g-actin. *J. Mol. Biol.* **230**, 186–195.

[36] Monasson, R. (1999). Diffusion,localization and dispersion relations on "small-world" lattices. *Eur. Phys. J.* **B 12**, 555–567.

[37] Biroli, G. & Monasson, R. (1999). A single defect approximation for localized states on random lattices. *J. Phys.* **A 32**, L255–L261.

[38] Anderson, P. W. (1958). Absence of diffusion in certain random lattices. *Physical Review,* **109**, 1492–1505.

[39] Anderson, P. W. (1978). Local moments and localized states. *Reviews of Modern Physics,* **50**, 191–201.

[40] Bahar, I., Atilgan, A. R., Jernigan, R. L. & Erman, B. (1997). Understanding the recognition of protein structural classes by amino acid composition. *Proteins: Structure Function and Genetics,* **29**, 172–185.

[41] Lustig, B., Bahar, I. & Jernigan, R. L. (1998). Rna bulge entropies in the unbound state correlate with peptide binding strengths for hiv–1 and biv tar rna because of improved conformational access. *Nucleic Acids Research,* **26**, 5212–5217.

[42] Bahar, I., Wallqvist, A., Covell, D. G. & Jernigan, R. L. (1998). Correlation between native–state hydrogen exchange and cooperative residue fluctuations from a simple model. *Biochemistry,* **39**, 1067–1075.

[43] Bahar, I. & Jernigan, R. L. (1998). Vibrational dynamics of transfer rnas: comparison of the free and synthetase–bound forms. *Journal of Molecular Biology,* **281**, 871–884.

[44] Jernigan, R. L., Demirel, M. C. & Bahar, I. (1999). Relating structure to function through the dominant modes of motion of dna topoisomerase ii. *International Journal of Quantum Chemistry,* **75**, 301–312.

[45] Bahar, I., Erman, B., Jernigan, R. L., Atilgan, A. R. & Covell, D. G. (1999). Collective motions in hiv–1 reverse transcriptase: examination of flexibility and enzyme function. *Journal of Molecular Biology,* **285**, 1023–1037.

[46] Haliloglu, T. & Bahar, I. (1999). Structure based analysis of protein dynamics. comparison of theoretical results for hen lysozyme with x–ray diffraction and nmr relaxation data. *Proteins,* **37**, 654–667.

[47] Bahar, I. & Jernigan, R. L. (1999). Cooperative fluctuations and subunit communication in tryptophan synthase. *Biochemistry,* **38**, 3478–3490.

[48] Keskin, O., Bahar, I. & Jernigan, R. L. (2000). Proteins with similar architectures exhibit similar large–scale dynamic behavior. *Biophysical Journal,* **78**, 2093–2106.

[49] Atilgan, A. R., Durell, S. R., Jernigan, R. L., Demirel, M. C., Keskin, O. & Bahar, I. (2001). Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical Journal,* **80**, 505–515.

[50] Cieplak, M. & Hoang, T. X. (2001). Kinetic non-optimality and vibrational stability of proteins. *Proteins: Structure Function and Genetics,* **44**, 20–25.

[51] Hoang, T. X. & Cieplak, M. (2000). Molecular dynamics of folding of secondary structures in go-type models of proteins. *J. Chem. Phys.* **112**, 6851–6862.

[52] Hoang, T. X. & Cieplak, M. (2000). Sequencing of folding events in go-type proteins. *J. Chem. Phys.* **113**, 8319–8328.

[53] Debe, D. A. & Goddard III, W. A. (1999). First principles prediction of protein folding rates. *J. Mol. Biol.* **294**, 619–625.

[54] Dinner, A. R. & Karplus, M. (2001). The roles of stability and contact order in determining protein folding rates. *Nat. Struct. Biol.* **8**, 21–22.

[55] Ivankov, D. N. & Finkelstein, A. (2001). Theoretical study of a landscape of protein folding-unfolding pathways. folding rates and midtransition. *Biochemistry,* **40**, 9957–9961.

[56] Erman, B. (2002). Determination of residue-residue interaction energy parameters and correlations from experimental thermal fluctuation data in proteins. *preprint,* .

[57] Diamond, R. (1990). On the use of normal modes in thermal parameter refinement: theory and application to the bovine pancreatic trypsin inhibitor. *Acta Cryst.* **A46**, 425–435.

[58] Doruker, P., Jernigan, R. L. & Bahar, I. (2002). Dynamics of large proteins through hierarchical levels of coarse-grained structures. *J. Comput. Chem.* **23**, 119–127.

[59] Halle, B. (2002). Flexibility and packing in proteins. *Proc. Natl. Acad. Sci. USA,* **99**, 1274–1279.

[60] Micheletti, C., Seno, F. & Maritan, A. (2000). Recurrent oligomers in proteins - an optimal scheme reconciling accurate and concise backbone representations in automated folding and design studies. *Proteins: Structure Function and Genetics,* **40**, 662–674.

[61] Bernstein, F. C., Koetzle, T. F., Williams, G. J., Jr, E. E. M., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The protein data bank: a computer-based archival file for macromolecular structures. *Journal of Molecular Biology,* **112**, 535–542.

[62] Condra et al., J. H. (1995). In-vivo emergence of hiv-1 variants resistant to multiple protease inhibitors. *Nature,* **374**, 569–571.

[63] Ala, P. J., Huston, E. E., Klabe, R. M., Jadhav, P. K. & and. C. H. Chang, P. Y. S. L. (1998). Counteracting HIV-1 protease drug resistance: structural analysis of mutant proteases complexed with XV638 and SD146, cyclic urea amides with broad specificities. *Biochemistry,* **37**, 15042–15049.

[64] Gulnik, S., Erickson, J. W. & Xie, D. (2000). Vitamins and hormones - advances in research and applications. *Vitam Horm.* **58**, 213–256.

[65] Wlodawer, A. & Erickson, J. W. (1993). Structure-based inhibitors of HIV-1 protease. *Annu Rev Biochem,* **62**, 543–585. and references therein.

[66] Reddy, P. & Ross, J. (1999). Amprenavir - A protease inhibitor for the treatment of patients with HIV-1 infection. *Formulary,* **34**, 567–675.

[67] Tisdale et al., M. (1995). Cross-resistance analysis of human-immunodeficiency-virus type-1 variants individually selected for resistance to 5 different protease inhibitors. *Antimicrob. Agents Chemother.* **39**, 1704–1710.

[68] Piana, S., CArloni, P. & Parrinello, M. (2002). Role of conformational fluctuations in the enzymatic reaction of hiv-1 protease. *J. Mol. Biol.* **in press**, ...

[69] Settanni, G., Hoang, T., Micheletti, C. & Maritan, A. (2002). Folding pathways of prion and doppel. *Biophysical Journal,* **in press**, in press.

[70] Lazaridis, T. & Karplus, M. (1997). "new view" of protein folding reconciled with the old through multiple unfolding simulations. *Science,* **278**, 1928–1931.